The Digital Universe –Information Theoretical Analyses

András Benczúr

Eötvös Loránd University

Department of Information Systems

Summary

Mankind gave born to a new universe, the Digital Universe. The majority of our data and information is inside it somewhere and in digital form of some kind. Even new observation – from digital sensors, cameras etc. – goes first in digital form into it.

It is huge. We know that a few zettabyte of information is collected in the Digital Universe. There are sophisticated procedures to measure how much information we have, but there is also an intensive discussion on what information is? Mathematical theories of information deal with quantitative properties. They mainly deal with the objective parts of information (representation and the mapping to their referents). The subjective aspect, the semantics of the referents is the problem of the observer. In [1] P.J. Denning summarizes the discussion on the definition of information in the following: *"The formal definitions of data (objective symbols) and information (subjective meaning) do not help me to design computers and algorithms. ... Still, what information is remains an open question. "*

If we want to get closer to the notions of information from the point of view of the mathematical models we have to investigate carefully what is measured by the entropy functions. According to Kolmogorov [2], we can measure the quantity of information in three ways.

All the three measures are related to the length of description and not to the meaning of information. They are connected to the length of optimal digital code. In the Shannon model, the expected value of the code length is minimized, whilst Kolmogorov entropy measures the minimal length of codes used by the selected optimal Universal Reference machine. In both models we do not know what information is, we only know that there is a way to construct/reconstruct it from a signal of given length. We do not know what information is, we only know how much it is.

The Digital Universe contains only the substitutions, or encodings of information, independently of whatever information means. Inside the Digital Universe the physical processes are either transformations of signals from one form to other one or they are materialized computations. So Digital Universe belongs to the territory of algorithmic information theory. The measure of the algorithmic information quantity, the Kolomogorov entropy is not good for the direct investigation of the Digital Universe. We explain that it is not the measure itself; but method of the selection and use of a Universal Reference Machine is important. We can use it as a measurement tool in finding approximation in quantitative analyses of the behavior of the Digital Universe.

Input devices of the Digital Universe form the new sensory systems of mankind. Our senses send superfluous information to our nervous system. Now the same is true for the Digital Universe. The balance between the collection and utilization of the sensory information is the result of the evolution for living individuals. Currently, one of the greatest challenges for computer science and informatics is to reach the balance with the Digital Universe. The Digital Universe is much more than a new communication system or a global automated information system as I discussed in [3].

Living in symbioses with the Digital Universe, we are observers of representational forms of information produced by computers that use both data and program codes stored on them. Entering new information – facts, data, queries - we always have to use representational forms that have been specified for computer acceptance. So we are in the realm of algorithmic information theory. This leads me to the hypothesis of the growing semantic gap between human beings and computers. With the growing of the size of databases the length of queries grows at least logarithmically, and may grow linearly. According to the estimation from IDC in [4] the size of the Digital Universe will grow in the next five year by a factor 9. It doubles every one and a half years. The most recent IDC study [5] projects that the digital universe will reach 40 zettabytes (ZB) by 2020, an amount that exceeds previous forecasts by 5 ZBs, resulting in a 50-fold growth from the beginning of 2010.

Paradoxically, inside the Digital Universe, the basic components, the physically existing – even temporarily - digits as bits and bytes have no semantic meaning but operational, computational or transformational. The observer's meanings at the very end of the interaction with the real world are in the mappings of the real world stuff to a formal computable model. This mapping is the kernel of filling the gap between human beings and computers.

We do not know whether the evolution of info-communication technologies will help us to fill the semantic gap, or the growing of the data volume stored on the Digital Universe will generate a growing semantic gap.

This talk uses algorithmic information theory to explain the problem of the growing semantic gap. What is important in algorithmic information theory is the role of the Universal Reference Machine. The Digital Universe is not only the largest collection of information but it is the anytime best approximation of the Universal Reference Machine. We gives a short introduction to algorithmic information theory.

We demonstrate in a simple model how the information quantity of queries and answers can be estimated under the condition that the content of the database is known. Also, we can estimate the effect of the size of the database on the length of queries and answers. We demonstrate this phenomenon by modeling the communication process between a computer, named Watson, and a person, named Holmes.

The effect of the growing data is demonstrated by some simple scenarios.

One weakness of algorithmic information theory is that it deals with the problem of "What": what is the data, what is the query and what is the answer. Then it estimates the sizes. Accessing the Digital Universe we must consider the other two dimensions of data: the Where and When dimensions. It is expressed by Alex Szalay [5] *"Data is everywhere, never be at a single location. Not scalable, not maintainable."*

At the end we analyze how this is related to the problem of Big Data and the advances in information harvesting.

References

[1] P.J. Denning: "What Have We Said About Computation? *Ubiquity Symposium, Closing Statement,* in *Ubiquity, an ACM Publication*, April, 2011. http://ubiquity.acm.org

[2], A.N. Kolmogorov, "Three approaches to the quantitative definition of information", *Problems of Information Transmission* **1** (1), 1{7, (1965).

[3] A. Benczúr: "The Evolution of Human Communication and the Information Revolution- A Mathematical Perspective", *Mathematical and Computer Modelling*, Vol. 38, No. 7-9. pp. 691-708. 2003.

[4] J. Gantz, D. Reinsel : "Extracting Value from Chaos" *International Data Corporation IVIEV* June 2011.

[5] J. Gantz ,D Reinsel: „THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East" International Data Corporation, December 2012, http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

[6] A. Szalay: „Extreme Data-Intensive Computing" , http://salsahpc.indiana.edu/tutorial/slides/0726/szalay-bigdata-2010.pdf